Improved Error Correction of NGS Data

29 de febrero de 2016

El trabajo realizado en el marco de esta tesis doctoral se centra en la corrección de errores en datos provenientes de técnicas de secuenciación masiva (también llamadas de nueva generación, Next Generation Sequencing o NGS) utilizando técnicas de computación intensiva.

Debido a la reducción de costes y el incremento en las prestaciones de los secuenciadores, así como en los avances en las ciencias médicas y biológicas, la cantidad de datos disponibles en NGS se ha incrementado notablemente. La utilización de computadores en el análisis de estas muestras se hace imprescindible para poder dar respuesta a la avalancha de información generada por estas técnicas. El uso de NGS transciende la investigación con numerosos ejemplos de uso clínico y agronómico (como por ejemplo la detección de mutaciones en tumores cancerosos), por lo que aparecen nuevas necesidades en cuanto al tiempo de proceso y la fiabilidad de los resultados. Para maximizar su aplicabilidad clínica, las técnicas de proceso de datos de NGS deben acelerarse y producir datos más precisos. En este contexto es en el que las técnicas de comptuación intensiva juegan un papel relevante. En la actualidad, es común disponer de computadores con varios núcleos de proceso e incluso utilizar múltiples computadores mediante técnicas de computación paralela distribuida, incluso fuera del uso científico. Las tendencias actuales hacia arquitecturas con un mayor número de núcleos (many-core) ponen de manifiesto que es ésta una aproximación relevante.

Esta tesis comienza con un análisis de los problemas fundamentales del proceso de datos en NGS de forma general y adaptado para su comprensión por una amplia audiencia, a través de una exhaustiva revisión del estado del arte en la corrección de datos de NGS. Esta revisión introduce gradualmente al lector en las técnicas de secuenciación masiva, presentando problemas y aplicaciones reales de las técnicas de NGS, destacando el impacto de esta tecnología en ciencia. De este estudio se concluyen dos ideas principales: La necesidad de analizar de forma adecuada las características de los datos de NGS, atendiendo a la enorme variedad intrínseca que tienen las diferentes técnicas de secuenciación masiva; y la necesidad de disponer de una herramienta versátil, eficiente y precisa para la corrección de errores, como fase previa a cualquier análisis genómico.

En el contexto del análisis de datos, la tesis presenta MuffinInfo. La herramienta MuffinInfo es una aplicación software implementada mediante HTML5 para favorecer su portabilidad tanto a nivel de sistema operativo como de dispositivo. MuffinInfo obtiene información relevante de datos crudos de NGS para

favorecer el entendimiento de sus características y la aplicación de técnicas de corrección de errores, soportando además la extensión mediante funciones que implementen estadísticos definidos por el usuario. MuffinInfo almacena los resultados del proceso en ficheros JSON que facilitan su integración en pipelines de proceso. Al usar HTML5, MuffinInfo puede funcionar en casi cualquier entorno hardware y software, dada el amplio soporte que tiene esta tecnología. La herramienta está implementada aprovechando múltiples hilos de ejecución y gestionando de forma concurrente el acceso a disco y la gestión del interfaz.

La segunda conclusión del análisis del estado del arte nos lleva a la oportunidad de aplicar de forma extensiva técnicas de computación de altas prestaciones en la corrección de errores para desarrollar una herramienta que soporte múltiples tecnologías (Illumina, Roche 454, Ion Torrent y experimentalmente PacBio). La herramienta propuesta (MuffinEC), soporta diferentes tipos de errores (sustituciones, delecciones, inserciones y valores desconocidos). MuffinEC supera los resultados obtenidos por las herramientas existentes en este ámbito, en los tres tipos de tests utilizados en la tesis. Ofrece una mejor tasa de corrección, en un tiempo muy inferior y utilizando menos recursos, lo que facilita además su aplicación en muestras de mayor tamaño en computadores convencionales, donde otras herramientas no pueden funcionar por problemas de recursos. MuffinEC utiliza una aproximación basada en etapas multiples. Primero agrupa todas las secuencias utilizando la métrica de los k-mers. En segundo lugar realiza un refinamiento de los grupos mediante el alineamiento con Smith-Waterman, generando contigs resultado del alineamiento múltiple de las secuencias compatibles en el grupo. Estos contigs resultan de la corrección por columnas de atendiendo a la frecuencia individual de cada base y la aplicación de diferentes fórmulas y técnicas que facilitan discriminar errores de variantes significativas.

La tesis se estructura por capítulos cuya base ha sido previamente publicada en revistas indexadas en posiciones destacadas del índice del Journal of Citation Reports y en congresos de prestigio.